

Applicant:

Masataka ANDOH, et al.

Title:

cDNA MICROARRAY DATA CORRECTION SYSTEM, METHOD,

PROGRAM, AND MEMORY MEDIUM

Appl. No.:

10/696,572

Filing Date: 10/30/2003

Examiner:

Unknown

Art Unit:

1651

CLAIM FOR CONVENTION PRIORITY

Commissioner for Patents P.O. Box 1450 Alexandria, VA 22313-1450

Sir:

The benefit of the filing date of the following prior foreign application filed in the following foreign country is hereby requested, and the right of priority provided in 35 U.S.C. § 119 is hereby claimed.

In support of this claim, filed herewith is a certified copy of said original foreign application:

> Japanese Patent Application No. 2003-124585 filed 04/28/2003.

> > Respectfully submitted,

Date: July 1, 2004

FOLEY & LARDNER LLP

Telephone:

Customer Number: 22428 (202) 672-5407

Facsimile:

(202) 672-5399

Articola 38,819 David A. Blumenthal

Attorney for Applicant

Registration No. 26,257

日本国特許庁 JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日 Date of Application:

2003年 4月28日

出願番号 Application Number:

特願2003-124585

[ST. 10/C]:

[JP2003-124585]

出 願 人
Applicant(s):

日本電気株式会社

大瀧 慈

社団法人バイオ産業情報化コンソーシアム

2003年 9月25日

特許庁長官 Commissioner, Japan Patent Office 今井康



【書類名】

特許願

【整理番号】

64002125

【特記事項】

特許法第30条第1項の規定の適用を受けようとする特

許出願

【提出日】

平成15年 4月28日

【あて先】

特許庁長官殿

【国際特許分類】

G06N 7/00

【発明者】

【住所又は居所】

東京都港区芝五丁目7番1号 日本電気株式会社内

【氏名】

安東 正貴

【発明者】

【住所又は居所】

東京都港区芝五丁目7番1号 日本電気株式会社内

【氏名】

斎藤 彰

【発明者】

【住所又は居所】 広島県廿日市市宮園 9 丁目 1 の 7

【氏名】

大瀧 慈

【発明者】

【住所又は居所】

広島県広島市佐伯区楽々園5-9 五日市住宅16-2

0 2

【氏名】

佐藤 健一

【発明者】

【住所又は居所】 広島県広島市南区仁保南2丁目8-7

【氏名】

西山 正彦 1

【発明者】

【住所又は居所】

東京都中央区八丁堀二丁目26番9号 グランデビルデ

ィング 社団法人バイオ産業情報化コンソーシアム内

【氏名】

大谷 敬子

【特許出願人】

【識別番号】

000004237

【氏名又は名称】 日本電気株式会社

【特許出願人】

【識別番号】

503077165

【氏名又は名称】 大瀧 慈

【特許出願人】

【識別番号】

500535301

【氏名又は名称】

社団法人バイオ産業情報化コンソーシアム

【代理人】

【識別番号】

100071272

【弁理士】

【氏名又は名称】

後藤 洋介

【選任した代理人】

【識別番号】

100077838

【弁理士】

【氏名又は名称】 池田 憲保

【手数料の表示】

【予納台帳番号】

012416

【納付金額】

21,000円

【提出物件の目録】

【物件名】

明細書 1

【物件名】

図面 1

【物件名】

要約書 1

【包括委任状番号】 0018587

【プルーフの要否】

要

【書類名】 明細書

【発明の名称】 c D N A マイクロアレイデータの補正システム、方法、プログラム、及び記録媒体

【特許請求の範囲】

【請求項1】 マイクロアレイデータのグローバル及びローカルな歪みに対してより精密な補正をし、さらに蛍光色素の感度の違いによる測定誤差を補正する c DNAマイクロアレイデータの補正システムにおいて、

各スポットのバックグラウンドノイズの除去及び信頼性を示すフラッグ情報を考慮し、あらかじめ調整されている遺伝子発現強度データを入力する入力装置と、前記入力された遺伝子発現強度データに対して、グリッド毎の順序統計量を用いて遺伝子発現強度データを規準化し、規準化された規準化遺伝子発現強度データを送出するデータ規準化手段と、前記規準化遺伝子発現強度データに対して、グリッドの座標におけるスポット位置に依存する歪みをノンパラメトリック平滑化法によって推定し、補正した第1の遺伝子発現強度補正データを送出するスポット位置による第1の補正手段と、前記第1の遺伝子発現強度補正データに対してS-D変換を行い、遺伝子発現強度データに潜在しうる蛍光色素の感度の違いによる歪みをノンパラメトリック平滑化法によって推定し、蛍光色素の感度の違いによる歪みが補正された第2の遺伝子発現強度補正データを送出する第2の補正手段とを具備するデータ解析装置と、

前記第2の遺伝子発現強度補正データを出力する出力装置を 有することを特徴とする c DNAマイクロアレイデータの補正システム。

【請求項2】 さらに、任意の段階で遺伝子発現強度データの歪みを定量化し、S-Dプロット上に視覚化するS-D変換手段を有していることを特徴とする請求項1記載のcDNAマイクロアレイデータの補正システム。

【請求項3】 前記順序統計量は以下の数1(尚、外1は、前記規準化遺伝子発現強度データであり、外2はチャンネルによって得られた全スポットの遺伝子発現強度データであり、 $L_k(c)$ および $M_k(c)$ はそれぞれグリッドkにおいて、チャンネルcによって得られた遺伝子発現強度データの25%点および50%点を示す。)で示されることを特徴とする請求項1又は2記載のcDNAマイ

クロアレイデータの補正システム。

【数1】

$$w_{\text{ij}}^{k}(c) = \frac{y_{\text{ij}}^{k}(c) - L_{k}(c)}{M_{k}(c) - L_{k}(c)}, \quad c = 1, 2, i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, K.$$

【外1】

 $w_{ij}^{k}(C)$

【外2】

y k (C)

【請求項4】 前記データ規準化手段は、少なくとも2つの遺伝子発現強度 データチャンネルによって得られた全スポットの遺伝子発現強度データを規準化 したかどうかを判定し、全スポットの遺伝子発現強度データを規準化するまで続けることを特徴とする請求項3記載のcDNAマイクロアレイデータの補正システム。

【請求項5】 前記規準化遺伝子発現強度データは、真の発現強度とスポット位置に依存する歪みとの和によって表されることを特徴とする請求項1記載のcDNAマイクロアレイデータの補正システム。

【請求項6】 前記第1の補正手段は、スポット位置に依存する歪みを x 軸 、 y 軸、及び前記 x , y 軸の交互作用による歪み(それぞれ、外 3 , 外 4 , 外 5 とする。)の回帰関係で示されるノンパラメトリック回帰モデルにより記述し、以下の数 2 に示されるノンパラメトリック平滑化法を用いて、スポット位置による歪み(外 6) を推定することを特徴とする請求項 1 記載の c DNAマイクロアレイデータの補正システム。

【数2】

 $\hat{\xi}_{\text{ij}}^{\,k}(c) = \hat{\alpha}_{\,k}^{\,(C)}\,(\text{i}) + \hat{\beta}_{\,k}^{\,(C)}\,(\text{j}) + \hat{\gamma}_{\,k}^{\,(C)}\,(\,(\,\text{i}-\text{m}_{\,\text{i}}\,)\,(\text{j}-\text{m}_{\,\text{j}}\,))\,\,,\,\,\,c=1,2,\,\,\text{i}=1,\cdots,\text{I}\,,\,\,\text{j}=1\,,\cdots,\text{J}\,.$

[外3]

 $\alpha_{\mathbf{k}}^{(C)}(\mathbf{i})$

【外4】

 $\beta_{k}^{(C)}(j)$

【外5】

 $\gamma_{\mathbf{k}}^{(C)}((i-m_i)(j-m_j))$ 【外 6】

 ξ_{ij}^{k} (C)

【請求項7】 前記スポット位置による歪みの補正は、以下の数3 (尚、外7は補正された真の遺伝子発現強度データである。) に従って行なわれることを特徴とする請求項6記載のcDNAマイクロアレイデータの補正システム。

【数3】

$$\hat{z}_{ij}^{k}(c) = w_{ij}^{k}(c) - \hat{\xi}_{ij}^{k}(c)$$

【外7】

îk (c)

【請求項8】 前記第2の補正手段における前記S-D変換は、以下の数4 に従って行なわれることを特徴とする請求項7記載のcDNAマイクロアレイデータの補正システム。

【数4】

$$u_{ij}^{k} = \hat{z}_{ij}^{k}(1) + \hat{z}_{ij}^{k}(2)$$

$$v_{ij}^{k} = \hat{z}_{ij}^{k}(1) - \hat{z}_{ij}^{k}(2)$$

【請求項9】 前記第2の補正手段は、以下の数5で示されるノンパラメトリック回帰モデルにより記述し、以下の数6及び数7で示されるノンパラメトリック平滑化法を用いて、蛍光色素の感度による測定誤差を推定し、補正を行うことを特徴とする請求項8記載のcDNAマイクロアレイデータの補正システム。

【数5】

$$v_{ij}^{k} = \phi(u_{ij}^{k}) + \epsilon_{ij}^{k}, \epsilon_{ij}^{k} = N(0, v^{2})$$

【数6】

$$\eta_{ij}^{k} = v_{ij}^{k} - \hat{\phi}(u_{ij}^{k})$$

【数7】

$$\hat{y}_{ij}^{k}(1) = \frac{1}{2} \left(u_{ij}^{k} + \eta_{ij}^{k} \right)$$

$$\hat{y}_{ij}^{k}(2) = \frac{1}{2} \left(u_{ij}^{k} - \eta_{ij}^{k} \right)$$

【請求項10】 前記補正の前提として、遺伝子が発現している確率は0. 5より小さいと仮定し、各グリッド内の半分以上のスポットで検出される蛍光強度は、バックグラウンドノイズあるいは系統誤差を示しているとすることを特徴とする請求項1記載のcDNAマイクロアレイデータの補正システム。

【請求項11】 さらに前記補正の前提として、グリッドにおいて、少なくとも2つの遺伝子発現強度データチャンネルによって得られた蛍光強度の25%点と50%点を、それぞれ $L_k(c)$ および $M_k(c)$ とするとき、遺伝子の大半は非発現状態にあり全てのグリッドとチャンネルにおいて蛍光強度の50%点以下の分布は共通であるという前提に基づき、 $L_k(c)$ と $M_k(c)$ - $L_k(c)$ は各グリッドおよび各チャンネルで等しいと仮定することを特徴とする請求項10記載のcDNAマイクロアレイデータの補正システム。

【請求項12】 マイクロアレイデータのグローバル及びローカルな歪みに

5/

対してより精密な補正をし、さらに蛍光色素の感度の違いによる測定誤差を補正する c DNAマイクロアレイデータの補正方法において、

各スポットのバックグラウンドノイズの除去及び信頼性を示すフラッグ情報を 考慮し、あらかじめ調整されている遺伝子発現強度データを入力するステップと 、 大半の遺伝子は発現していないことを前提として、前記入力された遺伝子発 現強度データに対してグリッド毎の順序統計量を用いて、当該遺伝子発現強度デ ータを規準化するステップと、

前記規準化された規準化遺伝子発現強度データを出力するステップと、

前記規準化遺伝子発現強度データに対して、グリッドの座標におけるスポット 位置に依存する歪みをノンパラメトリック平滑化法によって推定し、スポット位 置に依存したデータの歪みを補正するステップと、

前記スポット位置に依存したデータ歪みの補正がされた第1の遺伝子発現強度 補正データを出力するステップと、

前記第1の遺伝子発現強度補正データに対して、S-D変換を行い、遺伝子発現強度データに潜在しうる蛍光色素の感度の違いによる歪みをノンパラメトリック平滑化法によって推定し、蛍光色素の感度の違いによる歪みを補正するステップと、

前記蛍光色素の感度の違いによる歪みの補正がされた第2の遺伝子発現強度補 正データを出力するステップとを

有することを特徴とする c D N A マイクロアレイデータの補正方法。

【請求項13】 さらに、任意の段階で遺伝子発現強度データの歪みを定量化し、S-Dプロット上に視覚化するステップを有していることを特徴とする請求項12記載のcDNAマイクロアレイデータの補正方法。

【請求項14】 前記順序統計量は以下の数8(尚、外8は、前記規準化遺伝子発現強度データであり、外9はチャンネルによって得られた全スポットの遺伝子発現強度データであり、 $L_k(c)$ および $M_k(c)$ はそれぞれグリッドkにおいて、チャンネルcによって得られた遺伝子発現強度データの25%点および50%点を示す。)で示されることを特徴とする請求項12又は13記載のcDN Aマイクロアレイデータの補正方法。

【数8】

$$w_{ij}^{k}(c) = \frac{y_{ij}^{k}(c) - L_{k}(c)}{M_{k}(c) - L_{k}(c)}, \quad c = 1, 2, i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, K.$$

【外8】

 w_{ij}^{k} (C)

【外9】

y k (c)

【請求項15】 前記データを規準化するステップにおいて、少なくとも2つの遺伝子発現強度データチャンネルによって得られた全スポットの遺伝子発現強度データを規準化したかどうかを判定し、全スポットの遺伝子発現強度データを規準化するまで続けることを特徴とする請求項14記載のcDNAマイクロアレイデータの補正方法。

【請求項16】 前記規準化遺伝子発現強度データは、真の発現強度とスポット位置に依存する歪みとの和によって表されることを特徴とする請求項15記載のcDNAマイクロアレイデータの補正方法。

【請求項17】 前記スポット位置に依存したデータの歪みを補正するステップにおいて、スポット位置に依存する歪みを x 軸、 y 軸、及び前記 x , y 軸の交互作用による歪み(それぞれ、外10,外11,外12とする。)の回帰関係で示されるノンパラメトリック回帰モデルにより記述し、以下の数9で示されるノンパラメトリック平滑化法を用いて、スポット位置による歪み(外13)を推定することを特徴とする請求項12記載の c DNAマイクロアレイデータの補正方法。

【数9】

 $\hat{\xi}_{\text{ij}}^{k}(c) = \hat{\alpha}_{k}^{(C)}(\text{i}) + \hat{\beta}_{k}^{(C)}(\text{j}) + \hat{\gamma}_{k}^{(C)}((\text{i}-m_{\text{i}})(\text{j}-m_{\text{j}})) \text{ , } c=1,2,\text{ i=1,\cdots,I, j=1,\cdots,J} \text{ .}$

【外10】

 $\alpha_{\mathbf{k}}^{(C)}(\mathbf{i})$

【外11】

 $\beta_k^{(C)}(j)$

【外12】

$$\gamma_{\mathbf{k}}^{(C)}((i-m_i)(j-m_j))$$
 【外13】

 ξ_{ij}^{k} (C)

【請求項18】 前記スポット位置による歪みの補正は、以下の数10(尚、外14は補正された真の遺伝子発現強度データである。)に従って行なわれることを特徴とする請求項17記載のcDNAマイクロアレイデータの補正方法。

【数10】

$$\hat{z}_{ij}^{k}(c) = w_{ij}^{k}(c) - \hat{\xi}_{ij}^{k}(c)$$

【外14】

îk (C)

【請求項19】 前記蛍光色素の感度の違いによる歪みを補正するステップにおける前記S-D変換は、以下の数11に従って行なわれることを特徴とする請求項18記載のcDNAマイクロアレイデータの補正方法。

【数11】

$$u_{ij}^{k} = \hat{z}_{ij}^{k}(1) + \hat{z}_{ij}^{k}(2)$$

$$v_{ij}^{k} = \hat{z}_{ij}^{k}(1) - \hat{z}_{ij}^{k}(2)$$

【請求項20】 前記蛍光色素の感度の違いによる歪みを補正するステップにおいて、以下の数12で示されるノンパラメトリック回帰モデルにより記述し、以下の数13及び数14で示されるノンパラメトリック平滑化法を用いて、蛍光色素の感度による測定誤差を推定し、補正を行うことを特徴とする請求項19記載のcDNAマイクロアレイデータの補正方法。

【数12】

$$v_{ij}^{k} = \phi(u_{ij}^{k}) + \epsilon_{ij}^{k}$$
, $\epsilon_{ij}^{k} = N(0, v^{2})$

【数13】

$$\eta_{ij}^{k} = v_{ij}^{k} - \hat{\phi}(u_{ij}^{k})$$

【数14】

$$\hat{y}_{ij}^{k}(1) = \frac{1}{2} \left(u_{ij}^{k} + \eta_{ij}^{k} \right)$$

$$\hat{y}_{ij}^{k}(2) = \frac{1}{2} \left(u_{ij}^{k} - \eta_{ij}^{k} \right)$$

【請求項21】 前記補正の前提として、遺伝子が発現している確率は0. 5より小さいと仮定し、各グリッド内の半分以上のスポットで検出される蛍光強度は、バックグラウンドノイズあるいは系統誤差を示しているとすることを特徴とする請求項12記載のcDNAマイクロアレイデータの補正方法。

【請求項22】 さらに前記補正の前提として、グリッドにおいて、少なくとも2つの遺伝子発現強度データチャンネルによって得られた蛍光強度の25%点と50%点を、それぞれ $L_k(c)$ および $M_k(c)$ とするとき、遺伝子の大半は非発現状態にあり、全てのグリッドとチャンネルにおいて蛍光強度の50%点以下の分布は共通であるという前提に基づき、 $L_k(c)$ と $M_k(c)$ - $L_k(c)$ は各グリッドおよび各チャンネルで等しいと仮定することを特徴とする請求項21記載のcDNAマイクロアレイデータの補正方法。

【請求項23】 マイクロアレイデータのグローバル及びローカルな歪みに対してより精密な補正をし、さらに蛍光色素の感度の違いによる測定誤差を補正するためコンピュータに、

各スポットのバックグラウンドノイズの除去及び信頼性を示すフラッグ情報を 考慮し、あらかじめ調整されている遺伝子発現強度データを入力するステップと 、 大半の遺伝子は発現していないことを前提として、前記入力された遺伝子発 現強度データに対してグリッド毎の順序統計量を用いて、当該遺伝子発現強度デ ータを規準化するステップと、

前記規準化された規準化遺伝子発現強度データを出力するステップと、

前記規準化遺伝子発現強度データに対して、グリッドの座標におけるスポット 位置に依存する歪みをノンパラメトリック平滑化法によって推定し、スポット位 置に依存したデータの歪みを補正するステップと、

前記スポット位置に依存したデータ歪みの補正がされた第1の遺伝子発現強度 補正データを出力するステップと、

前記第1の遺伝子発現強度補正データに対して、S-D変換を行い、遺伝子発現強度データに潜在しうる蛍光色素の感度の違いによる歪みをノンパラメトリック平滑化法によって推定し、蛍光色素の感度の違いによる歪みを補正するステップと、

前記蛍光色素の感度の違いによる歪みの補正がされた第2の遺伝子発現強度補 正データを出力するステップとを

を実行させるための c DNAマイクロアレイデータ補正プログラム。

【請求項24】 マイクロアレイデータのグローバル及びローカルな歪みに対してより精密な補正をし、さらに蛍光色素の感度の違いによる測定誤差を補正するためコンピュータに、

各スポットのバックグラウンドノイズの除去及び信頼性を示すフラッグ情報を 考慮し、あらかじめ調整されている遺伝子発現強度データを入力するステップと 、 大半の遺伝子は発現していないことを前提として、前記入力された遺伝子発 現強度データに対してグリッド毎の順序統計量を用いて、当該遺伝子発現強度デ ータを規準化するステップと、 前記規準化された規準化遺伝子発現強度データを出力するステップと、

前記規準化遺伝子発現強度データに対して、グリッドの座標におけるスポット 位置に依存する歪みをノンパラメトリック平滑化法によって推定し、スポット位 置に依存したデータの歪みを補正するステップと、

前記スポット位置に依存したデータ歪みの補正がされた第1の遺伝子発現強度 補正データを出力するステップと、

前記第1の遺伝子発現強度補正データに対して、S-D変換を行い、遺伝子発現強度データに潜在しうる蛍光色素の感度の違いによる歪みをノンパラメトリック平滑化法によって推定し、蛍光色素の感度の違いによる歪みを補正するステップと、

前記蛍光色素の感度の違いによる歪みの補正がされた第2の遺伝子発現強度補 正データを出力するステップとを

を実行させるための c DNAマイクロアレイデータ補正プログラムを記録した コンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】

[0001]

【発明の属する技術分野】

本発明は、数理モデルに基づいた c DNAマイクロアレイデータのデータ補正システム、方法、プログラム及び記録媒体に関し、特にグローバルノーマライゼーションとローカルノーマライゼーション、さらに蛍光色素の感度の違いによる測定の歪みの補正をすることができる c DNAマイクロアレイデータの補正システム、方法、プログラム及び記録媒体に関するものである。

[0002]

【従来の技術】

現在、ゲノム研究は個々の遺伝子についての構造解析から体系的な遺伝子の機能解析へと展開しつつある。機能未知の遺伝子や総体としての遺伝子の機能解析のために、多数の遺伝子の発現強度を同時に定量化することのできる c DNA (相補的な DNA) マイクロアレイを用いた実験はその有効性が大いに期待されている。

[0003]

二色蛍光法による c DNAマイクロアレイを用いた実験の目的は二種類の細胞の遺伝子発現の違いを検出することにある。ここで、二色蛍光法による c DNAマイクロアレイの概要について述べる。まず、多数の遺伝子セットの c DNAを参照用のプローブとして、スライドグラス上にアレイ状に高密度に固定化する(マイクロアレイ)。

[0004]

次に、条件の異なる2種類のサンプル、細胞1と細胞2 (例えば正常細胞と癌細胞)から抽出したmRNAをそれぞれ波長の異なる蛍光色素でラベルし、ターゲットcDNAを合成する。そして、それらを等量混合したものをマイクロアレイに固定化された参照用のプローブcDNAに競合的にハイブリダイズさせる。ハイブリダイゼーション後、スキャナーでそれぞれの蛍光色素強度を測定する。細胞1にラベルされた蛍光色素をチャンネル1により、細胞2にラベルされた蛍光色素をチャンネル2により読み取り、それぞれを各細胞の遺伝子発現強度データ(マイクロアレイデータ)とする。

[0005]

このように、マイクロアレイデータが得られるまでの過程は複雑であり、高度な実験技術が必要とされることから、実験の各段階において様々な実験誤差が生じると考えられる。このため、マイクロアレイデータから真に生物学的意味のあるデータを取り出すためには遺伝子発現強度の分布と実験誤差の解析は解決すべき重要な課題である。

$[0\ 0\ 0\ 6\]$

遺伝子発現強度の分布に関しては、例えば、以下の非特許文献1を参照すると、Newton等は遺伝子発現強度にガンマ分布関数を仮定し、遺伝子発現強度比(チャンネル1とチャンネル2の遺伝子発現強度データの比)についての統計学的性質を考察している。

[0007]

また、観測された遺伝子発現強度データに対しては、例えば、以下の非特許文献2を参照すると、Lee等は真の遺伝子発現強度を2個の水準値に分離できる

ことおよび偶然誤差の存在を前提として、以下の数15に示されるような混合正 規分布を適用し、遺伝子発現強度データについての統計学的考察を行った。

【数15】

$$f(x) = p \phi(x - \mu_1 | \sigma_1^2) + (1-p)\phi(x - \mu_2 | \sigma_2^2)$$

ここで、x はスキャナーなどによって得られる蛍光強度などの遺伝子発現強度 データを表し、右辺第1項の外15は

【外15】

$$\phi(\mathbf{x} - \mu_1 | \sigma_1^2)$$

遺伝子が発現しているときの平均μ1、分散外16の正規分布、

【外16】

 σ_1^2

また、同第2項の外17は遺伝子が発現していないときの平均 μ 2、分散外18の正規分布の密度関数を表し、

【外17】

$$\phi(x - \mu_2 | \sigma_2^2)$$

【外18】

 σ_2^2

pはその混合率を表す母数である。

[0009]

実験誤差の解析については、系統誤差の除去、いわゆるノーマライゼーション

の方法がいくつか提案されている。ノーマライゼーションの方法は、大きく分けてアレイ上のすべてのスポットを対象にしたグローバルノーマライゼーションと,あるサブセットに分けた(例えばグリッド単位の)スポットを対象にしたローカルノーマライゼーションの二つが提案されている。グローバルノーマライゼーションについては、例えば、以下の非特許文献3を参照すると、Chen等は二つの細胞の遺伝子発現強度の中央値は等しいとしてチャンネル1とチャンネル2で得られた測定値の補正を行った。ローカルノーマライゼーションについては、例えば、以下の非特許文献4、5、6を参照すると、DudoitやSchuchhardtやYangは、系統誤差が、スポットのスライドグラス上の位置や、二種類の蛍光色素の感度の違いによって生じたものと考え、それらを除去する方法を提案した。

[0010]

【非特許文献1】

Newton et. al、2001年、ジャーナル・オブ・コンピュテーショナル・バイオロジー、第8巻、37~52頁(Journal of Computational Biology Vol. 8, pp. 37-52)

$[0\ 0\ 1\ 1]$

【非特許文献2】

Lee et. al、2000年、プロシーディング・オブ・ザ・ナショナル・アカデミィー・オブ・サイエンシズ、第97巻、第18号、9834~9839頁(Proceeding of the National Academy of Sciences Vol. 97, No 18, pp. 9834-9839)

$[0\ 0\ 1\ 2]$

【非特許文献3】

Chen et. al、1997年、ジャーナル・オブ・バイオメディカル・オプティクス、第2号、364~374頁(Journal of Biomedical Optics Vol. 2, pp. 364-374)

[0013]

【非特許文献4】

Dudoit et. al, 2000. Statistical methods for identifying differentia

lly expressed genes in replicated cDNA microarray experiments. Technical ~Report #578 2.

[0014]

【非特許文献5】

Schuchhardt et. al、2000年、ヌクレ・アシッド・リサーチ、第28巻、第10号 (Nucleic Acids Research, 2000, Vol.28, No. 10)

[0015]

【非特許文献6】

Yang et. al、 2002年、ヌクレ・アシッド・リサーチ、第30巻、第4号(Nuc leic Acids Research, 2002, Vol. 30, No. 4)

[0016]

【発明が解決しようとする課題】

上記した従来技術における問題点は、マイクロアレイデータの解析結果は再現性に乏しく不安定なものになりがちで、精度や効率は低いものとみなされていることである。その理由は、遺伝子の発現に関する真の信号と実験誤差の分離が十分に行われていないからである。その背景要因として、遺伝子の発現強度はそれぞれの遺伝子によってレベルが異なっていることが考えられ、その場合、上記数15によるモデルは明らかに過大に単純化されすぎている。

[0017]

本発明の目的は、マイクロアレイ上の遺伝子発現強度データに関してよりもっともな数理モデルを想定して、グローバルおよびローカルな歪みに対して高い精度の補正を行い、さらに蛍光色素の感度の違いによる測定誤差を補正するための包括的なノーマライゼーションの方法およびシステムを提供することである。

[0018]

【課題を解決するための手段】

本発明のcDNAマイクロアレイデータの補正システムは、蛍光強度などの遺伝子発現強度データを入力する入力装置と、プログラムの制御により動作するデータ解析装置と、出力装置とを含む。なお、入力される遺伝子発現強度データは、各スポットのバックグラウンドノイズの除去や各スポットの信頼性を示すフラ

ッグ情報を考慮し、あらかじめ調整されているものとする。

[0019]

前記データ解析装置は、下記の三個の連続した処理過程で構成される。第一処理過程であるデータ規準化手段では、前記入力装置から遺伝子発現強度データを入力し、大半の遺伝子は発現していないことを前提としてグリッド毎の順序統計量を用いて遺伝子発現強度データを規準化し、規準化した遺伝子発現強度データを出力する。

[0020]

第二処理過程であるスポット位置による補正手段では、前記規準化された遺伝子発現強度データを入力し、グリッド毎にスポット位置に依存する歪みをノンパラメトリック平滑化法によって推定し、スポット位置に依存したデータの歪みを補正した遺伝子発現強度データを出力する。

[0021]

第三処理過程であるS-Dプロットによる補正手段では、第二処理過程の段階まで補正された遺伝子発現強度データに対してMA変換の変形であるS-D変換(MA変換およびMAプロットについては、上記非特許文献6を参照)を行い、遺伝子発現強度データに潜在しうる蛍光色素の感度の違いによる歪みをノンパラメトリック平滑化法によって推定し、蛍光色素の感度の違いによる歪みを補正した遺伝子発現強度データを前記出力装置に出力する。

[0022]

なお、本システムは、任意の段階で遺伝子発現強度データの歪みを定量化し、 S-Dプロット上に視覚化するS-D変換手段を有していることを特徴とする。

[0023]

このような構成を採用し、遺伝子発現強度データを補正することにより、本発明の目的を達成することができる。

[0024]

【発明の実施の形態】

はじめに、本発明におけるマイクロアレイの構造を説明する。図1を参照すると、K個の各グリッドにI×J個ずつ、合計K×I×J個のcDNAがスライド

グラス上にスポットされている。いま、グリッドkにおける座標(i, j)にスポットされたcDNAに対して、チャンネルc=1, 2によって得られた蛍光強度を外19とする。

[0025]

【外19】

 $y_{ij}^{k}(c), c = 1,2$

次に、以下の2つの仮定をする。

[0026]

(仮定1)

遺伝子が発現している確率は0.5より小さいと仮定し、各グリッド内の半分以上のスポットで検出される蛍光強度外20は、バックグラウンドノイズあるいは系統誤差を示しているとする。

[0027]

【外20】

y k (c)

(仮定2)

グリッドkにおいて、チャンネル c によって得られた蛍光強度外 2 1 の 2 5 % 点と 5 0 % 点を、

【外21】

y k (c)

それぞれ $L_k(c)$ および $M_k(c)$ とするとき、遺伝子の大半は非発現状態にあり全てのグリッドとチャンネルにおいて蛍光強度の 50% 点以下の分布は共通であるという前提に基づき、 $L_k(c)$ と $M_k(c)$ ー $L_k(c)$ は各グリッドおよび各チャンネルで等しいと仮定する。

[0028]

次に、以上の仮定をもとに、本発明の第1の実施の形態について図面を参照して詳細に説明する。図2を参照すると、本発明の第1の実施の形態は、蛍光強度などの遺伝子発現強度データを入力する入力装置1と、プログラム制御により動作するデータ解析装置2と、ディスプレイ装置や印刷装置等の出力装置3とを含む。データ解析装置は、データ規準化手段21と、スポット位置による補正手段22と、S-Dプロットによる補正手段23とを備えている。

[0029]

データ規準化手段21は、与えられた遺伝子発現強度データに対して、グリッド毎の順序統計量を用いて遺伝子発現強度データを規準化し、スポット位置による補正手段22及びS-D変換手段24に送る。

[0030]

スポット位置による補正手段22は、データ規準化手段21から送られてきた 規準化された遺伝子発現強度データに対して、グリッド毎にスポット位置に依存 する歪みをノンパラメトリック平滑化法によって推定し、補正した遺伝子発現強 度データをS-Dプロットによる補正手段23及びS-D変換手段24に送る。

[0031]

S-Dプロットによる補正手段23は、スポット位置による補正手段22から送られてきた補正された遺伝子発現強度データにS-D変換を行い、蛍光色素の感度の違いに起因する歪みをノンパラメトリック平滑化法により補正した後、遺伝子発現強度データを出力装置3へ送る。

[0032]

S-D変換手段24は送られてきた遺伝子発現強度データにS-D変換を行い、出力装置3へ送る。

[0033]

次に、図2、図3を参照して本実施の形態について詳細に説明する。入力装置 1より入力された蛍光強度などの遺伝子発現強度データはデータ規準化手段21 へ送られる。データ規準化手段21は、送られてきた発現強度データに対して、 以下の数16で示されるように、グリッド毎の順序統計量を用いて発現強度デー タを規準化する(図3のステップA1)。

[0034]

【数16】

$$w_{ij}^{k}(c) = \frac{y_{ij}^{k}(c) - L_{k}(c)}{M_{k}(c) - L_{k}(c)}, \quad c = 1, 2, i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, K.$$

2つのチャンネルによって得られた全スポットの遺伝子発現強度データ外 2 2 を規準化したかどうかを判定し、

【外22】

y^kij (c)

全スポットの遺伝子発現強度データ($2 \times I \times J \times K$ 個)を規準化するまで続ける(ステップA2)。

[0035]

データ規準化手段21において規準化された遺伝子発現強度データ外23に対して、

【外23】

 $w_{ij}^{k}(C)$

外24を真の発現強度を反映した蛍光強度(以下、真の発現蛍光強度)とし、

【外24】

 $z_{ij}^{k}(C)$

外 2 5 を グリッド k の 座標 (i, j) に おけるスポット 位置に 依存する 歪みとする。

[0036]

【外25】

 ξ_{ij}^{k} (C)

このとき、以下の数17に示すように、遺伝子発現強度データ外26は、真の発現強度外27とスポット位置に依存する歪み外28との和によって表されるとする。

[0037]

【外26】

 $w_{ij}^{k}(c)$

【外27】

 $z_{ij}^{k}(c)$

【外28】

 ξ_{ij}^{k} (c)

【数17】

 $w_{ij}^{k}(c) = z_{ij}^{k}(c) + \xi_{ij}^{k}(c) + \varepsilon_{ij}^{k}(c), \quad \varepsilon_{ij}^{k}(c) \sim N(0, \sigma_{k}(c)^{2}), \quad c = 1, 2.$

ただし、外29はランダムなノイズであるとする。

[0038]

【外29】

 ϵ_{ij}^{k} (C)

スポット位置による補正手段22は、以下の数18に示すようにスポット位置

に依存する歪み外30を「x軸」、「y軸」および「2つの軸の交互作用」による歪みの回帰関係で示されるノンパラメトリック回帰モデルにより記述し、

【外30】

 ξ_{ij}^{k} (C)

以下の数19に示すようにノンパラメトリック平滑化法を用いて、スポット位置 による歪み外31を推定する。

[0039]

【外31】

 ξ_{ij}^{k} (C)

【数18】

 $\xi_{ij}^{k}(c) = \alpha_{k}^{(C)}(i) + \beta_{k}^{(C)}(j) + \gamma_{k}^{(C)}((i - m_{i})(j - m_{j})), c=1,2, i=1,\cdots,I, j=1,\cdots,J,$

 $\sum_{\bf i} \, \alpha_{\bf k}^{(C)}({\bf i}) = 0 \, , \, \sum_{\bf j} \beta_{\bf k}^{(C)}({\bf j}) = 0 \, , \, \sum_{\bf i} \, \sum_{\bf j} \gamma_{\bf k}^{(C)}(({\bf i} - m_{\bf i})({\bf j} - m_{\bf j})) = 0 \, . \label{eq:constraint}$

【数19】

 $\hat{\xi}_{\text{ij}}^{k}(\text{c}) = \hat{\alpha}_{k}^{(\text{C})}(\text{i}) + \hat{\beta}_{k}^{(\text{C})}(\text{j}) + \hat{\gamma}_{k}^{(\text{C})}\left((\text{i}-\text{m}_{\text{i}})(\text{j}-\text{m}_{\text{j}})\right) \;,\; \text{C=1,2, i=1,\cdots,I, j=1,\cdots,J} \;.$

ここで、外32とする。

[0040]

【外32】

 $m_i = \lfloor I / 2 \rfloor$, $m_j = \lfloor J / 2 \rfloor$

外33はα以上の最小の整数とする。

[0041]

【外33】

スポット位置による補正手段22は、以下の数20に示すように、データ規準 化手段21において規準化された遺伝子発現強度データ外34に対して、推定さ れたスポット位置による歪み外35を補正する(ステップA3)。

[0042]

【外34】

 $w_{ij}^{k}(c)$

【外35】

 $\hat{\xi}_{ij}^{k}$ (c)

【数20】

$$\hat{z}_{ij}^{k}(c) = w_{ij}^{k}(c) - \hat{\xi}_{ij}^{k}(c)$$

データ規準化手段21において規準化された全スポットの遺伝子発現強度データ外36に対して、

【外36】

 w_{ij}^{k} (C)

. スポット位置による歪み外37の補正をしたかどうかを判定し、

【外37】

 $\hat{\xi}_{ij}^{k}$ (C)

全スポットの遺伝子発現強度データ($2 \times I \times J \times K$ 個)を補正するまで続ける(ステップA4)。

[0043]

S-Dプロットによる補正手段23は、スポット位置による補正手段22において補正された真の遺伝子発現強度データ外38に対して、

【外38】

îk (C)

以下の数21に示すように、S-D変換を行う。

[0044]

【数21】

$$u_{ij}^{k} = \hat{z}_{ij}^{k}(1) + \hat{z}_{ij}^{k}(2)$$

$$v_{ij}^{k} = \hat{z}_{ij}^{k}(1) - \hat{z}_{ij}^{k}(2)$$

さらに、以下の数22で示されるようなノンパラメトリック回帰モデルを記述 し、以下の数23及び数24に示すようにノンパラメトリック平滑化法を用いて 蛍光色素の感度による測定誤差を推定し、補正を行う(ステップA5)。

[0045]

【数22】

$$v_{ij}^{k} = \phi(u_{ij}^{k}) + \epsilon_{ij}^{k}$$
 , $\epsilon_{ij}^{k} = N(0, v^{2})$

【数23】

$$\eta_{ij}^{k} = v_{ij}^{k} - \hat{\phi}(u_{ij}^{k})$$

【数24】

$$\hat{y}_{ij}^{k}(1) = \frac{1}{2} \left(u_{ij}^{k} + \eta_{ij}^{k} \right)$$

$$\hat{y}_{ij}^{k}(2) = \frac{1}{2} \left(u_{ij}^{k} - \eta_{ij}^{k} \right)$$

スポット位置による補正手段22において補正された真の遺伝子発現強度データ外39に対して、

【外39】

 \hat{z}_{ij}^{k} (c)

S-Dプロットによる補正をしたかどうかを判定し、全スポットの真の遺伝子発現強度データ(2×I×J×K個)を補正するまで続ける(ステップA6)。

[0046]

なお、図3のA2、A4の各ステップ終了後、遺伝子発現強度データはS-D変換手段24を介して出力装置3に送られ、S-Dプロットによって遺伝子発現強度データの歪みを視覚化することができる。

[0047]

次に、本実施の形態の効果について説明する。本実施の形態では、グリッド間での順序統計量を用いた規準化(グローバルノーマライゼーション)とグリッド内でのスポット位置に依存する歪みの補正(ローカルノーマライゼーション)を組み合わせたノーマライゼーションを行った。これにより、グリッド間での遺伝子発現強度の偏りによる系統誤差と、グリッド内でのスポット位置に依存する歪みを同時に補正することができる。さらに、S-Dプロットによる補正においては、発現強度データの和と差を用いることにより、蛍光色素の感度の違いによる測定誤差を補正することができる。

[0 0 4 8]

次に、本発明の第2の実施の形態について図面を参照して詳細に説明する。図4を参照すると、本発明の第2の実施の形態は、本発明の第1の実施の形態と同様に、入力装置、データ解析装置、出力装置を備え、更に、データ解析プログラムを記録した記録媒体4を備える。この記録媒体4は可搬形あるいは固定型のいずれであってもよく、磁気ディスク、半導体メモリ、CD-ROMその他の記録媒体であってもよい。

[0049]

また、本手法を実行できるコンピュータプログラムを、ネットワークに接続されたコンピュータの記録装置に格納しておき、ネットワークを介して他のコンピュータに転送することもできる。本アルゴリズムを実行するコンピュータプログラムを提供する提供媒体としては、様々な形式のコンピュータに読み出し可能な媒体として頒布可能であって、特定のタイプの媒体に限定されるものではない。データ解析プログラムは記録媒体4からデータ解析装置5に読み込まれ、データ

解析装置2の動作を制御し、入力装置1から入力されたデータファイルに対して 第1の実施の形態におけるデータ処理装置2による処理と同一の処理を実行する

[0050]

【実施例】

以下、本発明の実施例について説明する。例として用いたデータは、異なる2種類の癌細胞(A細胞、B細胞)の遺伝子発現状況の比較のために行われた実験から得られたものである。

[0051]

一枚のチップ上に48グリッド,1グリッドあたり441(21×21)スポット,計21168の遺伝子の発現パターンについて調べたものである。

[0052]

図5、図7はチャンネル1により得られたオリジナルデータのA細胞遺伝子発現強度を示し、図6、図8はチャンネル2によって得られたオリジナルデータのB細胞遺伝子発現強度を示す。それぞれの図は、マイクロアレイ上のスポット位置に対する遺伝子発現強度の対数値をプロットしたものである。また、図7、図8は第1グリッドから第4グリッドまでを拡大したものである。図5~図8を見ると、遺伝子発現強度がグリッドごとに周期的に繰り返される系統的な歪みが観察される。マイクロアレイ上の遺伝子は無作為にスポットされているので、このような歪みは実験誤差と考えられる。

[0053]

図9は、それらのS-Dプロットである。横軸は、各チャンネルの遺伝子発現強度の和、縦軸はそれらの差をとったものを示している。各チャンネルの遺伝子発現強度の和が小さい領域と大きい領域においては、各チャンネルの遺伝子発現強度の差は真の遺伝子発現の違いによる影響は小さく、各チャンネルの蛍光色素の感度の違いによるものと考えられる。これにより、図9において蛍光色素の感度の違いによって生じたと考えられる歪みが観察される。

[0054]

図10に、チャンネル1におけるオリジナルデータのスポット位置に対する遺

伝子発現強度の図を示す。図11に、チャンネル1における第一処理過程後のスポット位置に対する遺伝子発現強度の図を示す。図12に、チャンネル1における第二処理過程後のスポット位置に対する遺伝子発現強度の図を示す。スポット位置に依存していたグリッドごとに周期的に繰り返される系統的な歪みが補正されて取り除かれていることがわかる。

[0055]

図13にチャンネル1における第三処理過程後のスポット位置に対する遺伝子発現強度の図を示す。図14~図17にチャンネル2におけるオリジナルデータ、第一処理過程後、第二処理過程後、第三処理過程後のスポット位置に対する遺伝子発現強度の図を示す。チャンネル1と同様にスポット位置に依存していたグリッドごとに周期的に繰り返される系統的な歪みが補正されて取り除かれていることがわかる。

[0056]

図18~図21にオリジナルデータ、第一処理過程後、第二処理過程後、第三処理過程後のS-Dプロットを示す。図21を見ると、蛍光色素の感度の違いによる歪みが補正されて取り除かれていることがわかる。

[0057]

【発明の効果】

本発明によれば、グリッド間での位置および尺度の揺らぎに対する頑健な順序統計量の25%点と50%点による規準化(グローバルノーマライゼーション)と、グリッド内でのスポット位置に依存する歪みの補正(ローカルノーマライゼーション)を組み合わせてノーマライゼーションを行っているため、グリッド間での遺伝子発現強度の偏りや感度の揺らぎによる系統誤差と、グリッド内でのスポット位置に依存する歪みを、発現している遺伝子の頻度や外れ値の影響をほとんど受けることなく同時に補正することができる。

$[0\ 0\ 5\ 8]$

又、本発明によれば、S-Dプロットにおいて遺伝子発現強度データの和と差を用いることによって、それぞれの蛍光色素の感度の違いが得られ易く、それによる測定誤差を的確に抽出することができるため、蛍光色素の感度の違いによる

測定の歪みを効率良く補正することができる。

【図面の簡単な説明】

図1

本発明におけるマイクロアレイの構造を示す図である。

[図2]

本発明の第1の実施の形態の構成を示すブロック図である。

【図3】

本発明の第1の実施の形態の動作を示す流れ図である。

図4

本発明の第2の実施の形態の構成を示すブロック図である。

【図5】

チャンネル1で得られたオリジナルデータの遺伝子発現強度の図である。

図6

チャンネル2で得られたオリジナルデータの遺伝子発現強度の図である。

【図7】

チャンネル1で得られたオリジナルデータ(第1グリッドから第4グリッド) の遺伝子発現強度の図である。

【図8】

チャンネル2で得られたオリジナルデータ(第1グリッドから第4グリッド) の遺伝子発現強度の図である。

【図9】

オリジナルデータに対するS-Dプロットである。

【図10】

チャンネル1のオリジナルデータの遺伝子発現強度の図である。

【図11】

チャンネル1の第一処理過程後の遺伝子発現強度の図である。

【図12】

チャンネル1の第二処理過程後の遺伝子発現強度の図である。

【図13】

チャンネル1の第三処理過程後の遺伝子発現強度の図である。

図14

チャンネル2のオリジナルデータの遺伝子発現強度の図である。

【図15】

チャンネル2の第一処理過程後の遺伝子発現強度の図である。

【図16】

チャンネル2の第二処理過程後の遺伝子発現強度の図である。

【図17】

チャンネル2の第三処理過程後の遺伝子発現強度の図である。

【図18】

オリジナルデータに対するS-Dプロットである。

【図19】

第一処理過程後のS-Dプロットである。

【図20】

第二処理過程後のS-Dプロットである。

【図21】

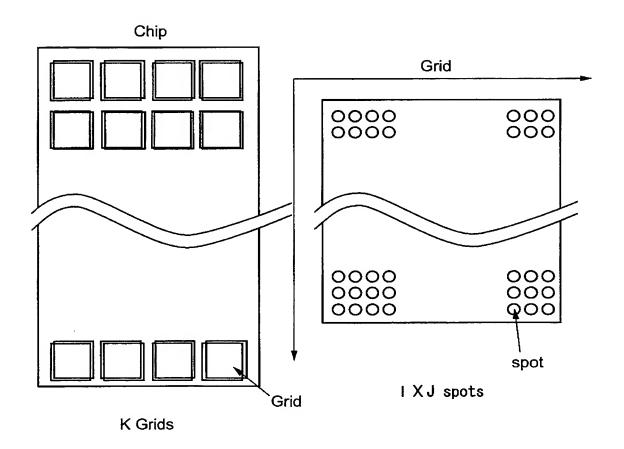
第三処理過程後のS-Dプロットである。

【符号の説明】

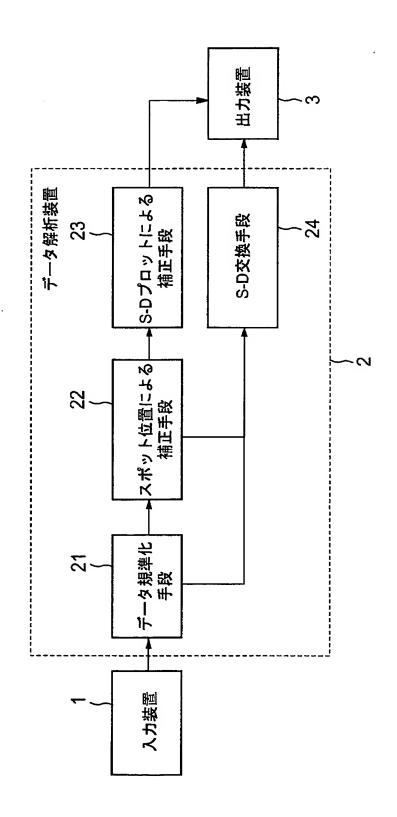
- 1 入力装置
- 2 データ解析装置
- 3 出力装置
- 21 データ規準化手段
- 22 スポット位置による補正手段
- 23 S-Dプロットによる補正手段
- 24 S-D変換手段

【書類名】 図面

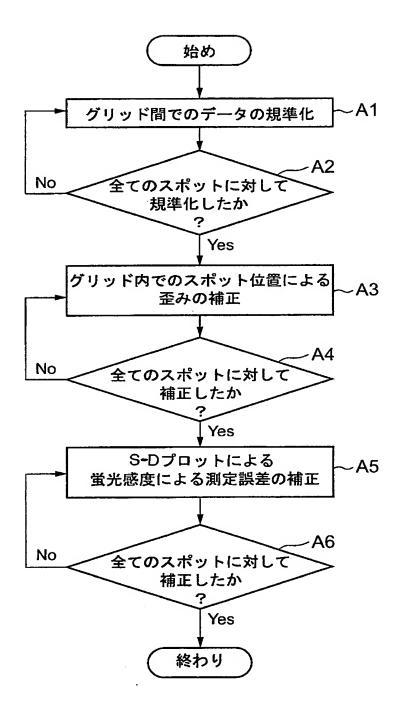
図1】



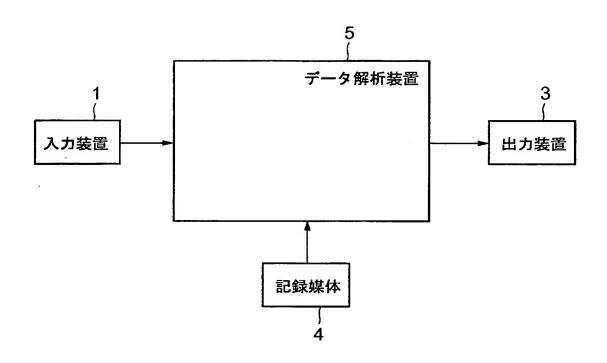
[図2]



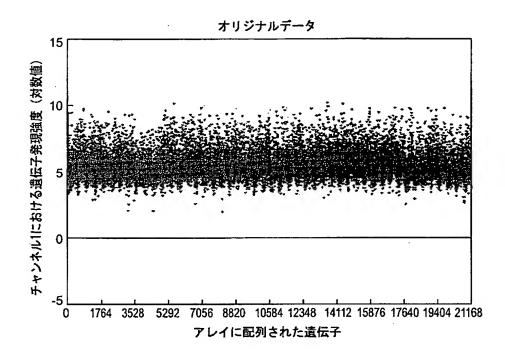
【図3】



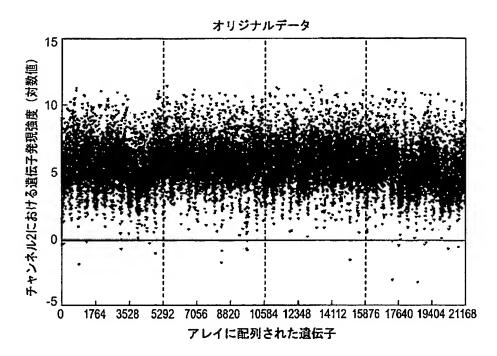
【図4】



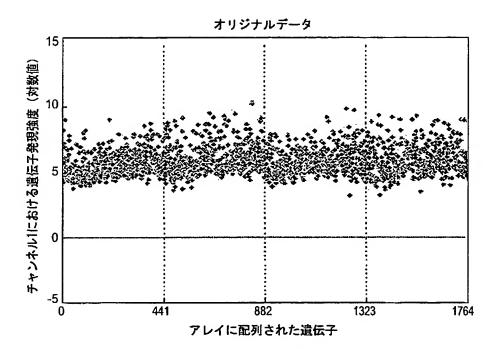
【図5】



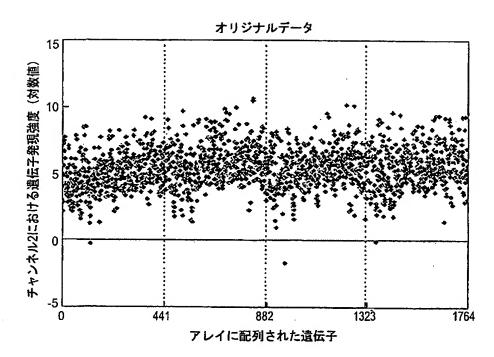
【図6】



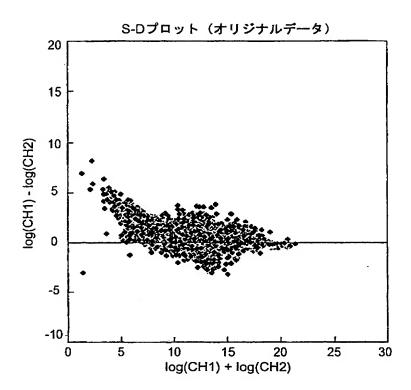
7/



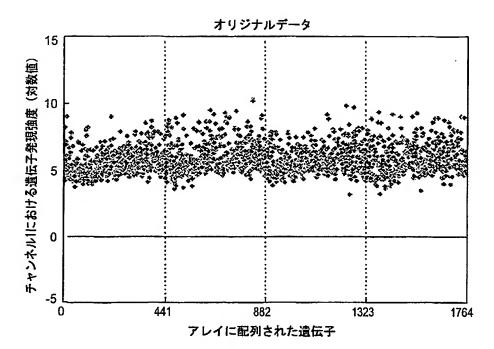
【図8】



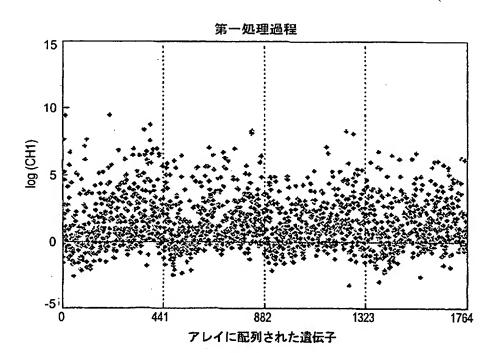
【図9】



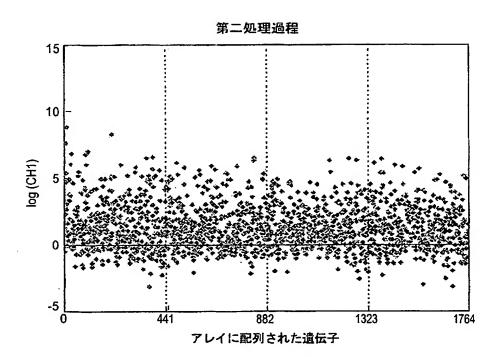
【図10】



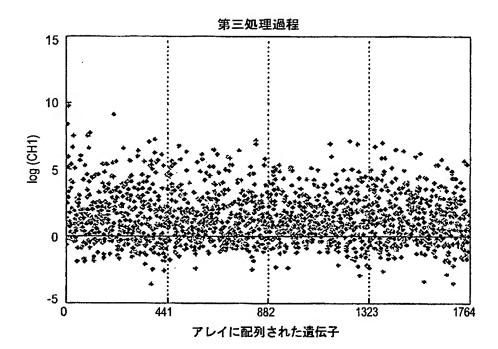
【図11】



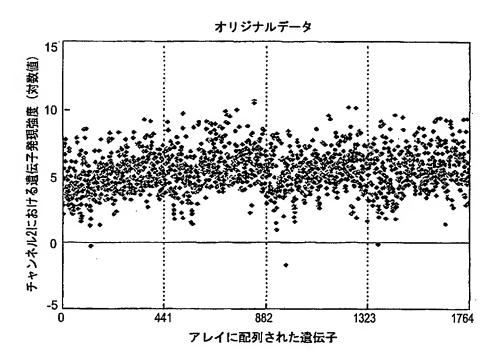
【図12】



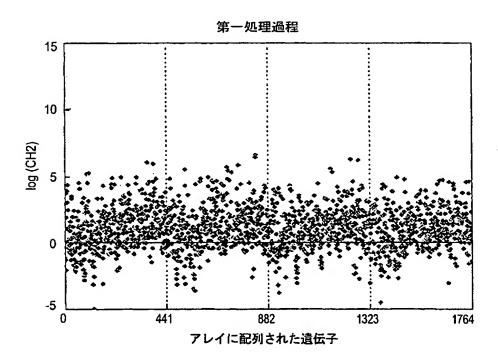
【図13】



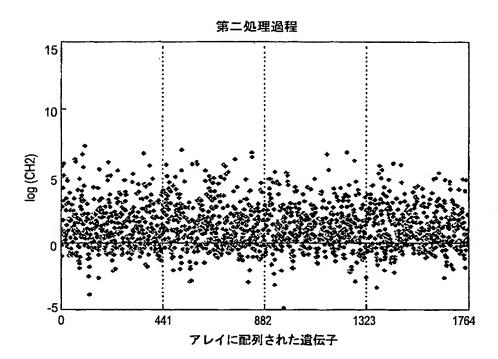
【図14】



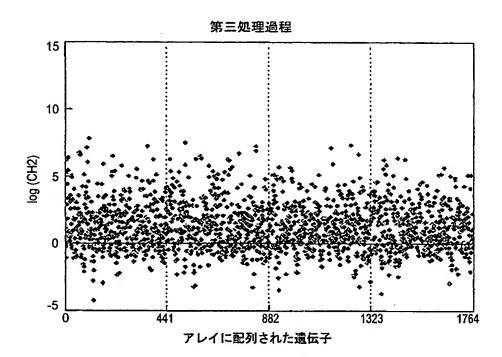
【図15】



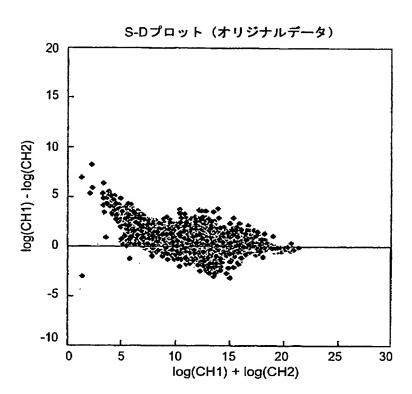
【図16】



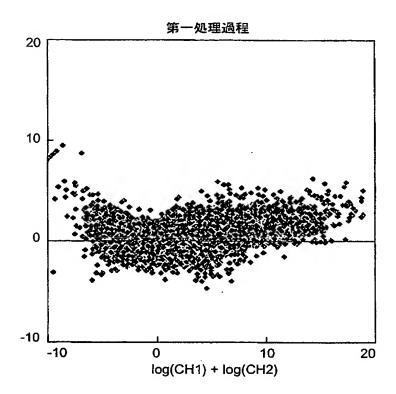
【図17】



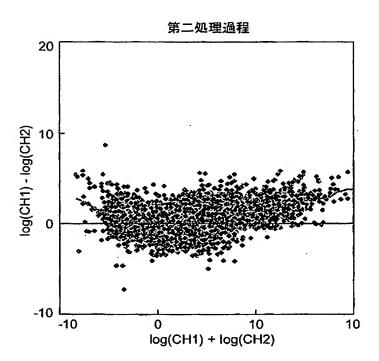
【図18】



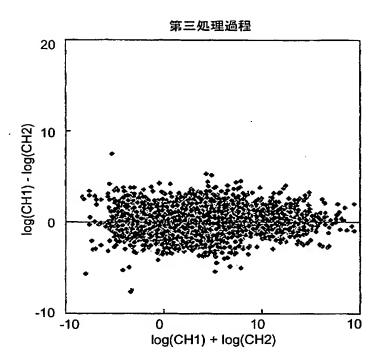
【図19】



【図20】



【図21】



1/E

【書類名】 要約書

【要約】

【課題】 マイクロアレイデータのグローバルおよびローカルな歪みに対してより精密な補正をし、さらに蛍光色素の感度の違いによる測定誤差を補正する。

【解決手段】 第一処理過程であるデータ規準化手段は、入力装置から遺伝子発現強度データを入力し、大半の遺伝子は発現していないことを前提としてグリッド毎の順序統計量を用いて遺伝子発現強度データを規準化し、規準化した遺伝子発現強度データを出力する。第二処理過程であるスポット位置による補正手段は、グリッド毎にスポット位置に依存する歪みをノンパラメトリック平滑化法によって推定し、スポット位置に依存した歪みを補正した遺伝子発現強度データを出力する。第三処理過程であるS-Dプロットによる補正手段は、S-D変換を行い、蛍光色素の感度の違いによる歪みをノンパラメトリック平滑化法によって推定し、蛍光色素の感度の違いによる歪みを補正した遺伝子発現強度データを出力装置に出力する。

【選択図】 図2

特願2003-124585

出願人履歴情報

識別番号

[000004237]

1. 変更年月日 [変更理由] 1990年 8月29日

住 所

新規登録

東京都港区芝五丁目7番1号

氏 名 日本電気株式会社

特願2003-124585

出願人履歴情報

識別番号

[503077165]

1. 変更年月日 [変更理由]

2003年 2月26日

新規登録

住 所

広島県廿日市市宮園9丁目1の7

氏 名 大瀧 慈



出願人履歴情報

識別番号

[500535301]

1. 変更年月日

2000年11月20日

[変更理由]

新規登録

住所

東京都中央区八丁堀二丁目26番9号 グランデビルディング

氏 名 社団法人バイオ産業情報化コンソーシアム